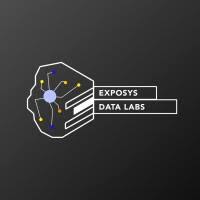
EXPOSYS DATA LABS

YELAHANKA, BENGALURU - 560064

****

**INTERNSHIP**

**PROFIT PREDICTION BY USING VARIOUS REGRESSION MODELS**

*Submitted By*

**Puja S**

*Under the guidance of*

Mr. Aravind Kumar

R&D Engineer

2022-2023

**ABSTRACT**

Data science is an [interdisciplinary](https://en.wikipedia.org/wiki/Interdisciplinary) field that uses [scientific methods](https://en.wikipedia.org/wiki/Scientific_method), processes, [algorithms](https://en.wikipedia.org/wiki/Algorithm) and systems to extract or extrapolate [knowledge](https://en.wikipedia.org/wiki/Knowledge) and insights from noisy, structured and [unstructured data](https://en.wikipedia.org/wiki/Unstructured_data), and apply knowledge from data across a broad range of application domains. Data science is related to [data mining](https://en.wikipedia.org/wiki/Data_mining), [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and [big data](https://en.wikipedia.org/wiki/Big_data).

Data science is a "concept to unify [statistics](https://en.wikipedia.org/wiki/Statistics), [data analysis](https://en.wikipedia.org/wiki/Data_analysis), [informatics](https://en.wikipedia.org/wiki/Informatics), and their related [methods](https://en.wikipedia.org/wiki/Scientific_method)" in order to "understand and analyse actual [phenomena](https://en.wikipedia.org/wiki/Phenomena)" with [data](https://en.wikipedia.org/wiki/Data). It uses techniques and theories drawn from many fields within the context of [mathematics](https://en.wikipedia.org/wiki/Mathematics), statistics, [computer science](https://en.wikipedia.org/wiki/Computer_science), [information science](https://en.wikipedia.org/wiki/Information_science), and [domain knowledge](https://en.wikipedia.org/wiki/Domain_knowledge). However, data science is different from [computer science](https://en.wikipedia.org/wiki/Computer_science) and information science. [Turing Award](https://en.wikipedia.org/wiki/Turing_Award) winner [Jim Gray](https://en.wikipedia.org/wiki/Jim_Gray_(computer_scientist)) imagined data science as a "fourth paradigm" of science ([empirical](https://en.wikipedia.org/wiki/Empirical_research), [theoretical](https://en.wikipedia.org/wiki/Basic_research), [computational](https://en.wikipedia.org/wiki/Computational_science), and now data-driven) and asserted that "everything about science is changing because of the impact of [information technology](https://en.wikipedia.org/wiki/Information_technology)" and the [data deluge](https://en.wikipedia.org/wiki/Information_explosion).

A data scientist is someone who creates programming code and combines it with statistical knowledge to create insights from data.

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **SL NO.** | **DESCRIPTION** | **PAGE NO.** |
| CHAPTER 1 | **Introduction** | 4 |
| CHAPTER 2 | **Existing Method** | 10 |
| CHAPTER 3 | **Proposed Method** | 12 |
| CHAPTER 4 | **Methodology** | 13 |
| CHAPTER 6 | **Results** | 15 |
| CHAPTER 7 | **Conclusion** | 19 |
|  | **References** | 20 |

**CHAPTER 1**

**INTRODUCTION**

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Also called simple regression or ordinary least squares (OLS), linear regression is the most common form of this technique. Linear regression establishes the [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between two variables based on a [line of best fit](https://www.investopedia.com/terms/l/line-of-best-fit.asp). Linear regression is thus graphically depicted using a straight line with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of one variable when the value of the other is zero. [Non-linear regression](https://www.investopedia.com/terms/n/nonlinear-regression.asp) models also exist, but are far more complex.

Regression analysis is a powerful tool for uncovering the associations between variables observed in data, but cannot easily indicate causation. It is used in several contexts in business, finance, and economics. For instance, it is used to help investment managers value assets and understand the relationships between factors such as [commodity prices](https://www.investopedia.com/terms/c/commodity.asp) and the stocks of businesses dealing in those commodities.

Regression as a statistical technique should not be confused with the concept of regression to the mean ([mean reversion](https://www.investopedia.com/terms/m/meanreversion.asp)).

**Types of Regression Analysis Techniques**

There are many types of regression analysis techniques, and the use of each method depends upon the number of factors. These factors include the type of target variable, shape of the regression line, and the number of independent variables.

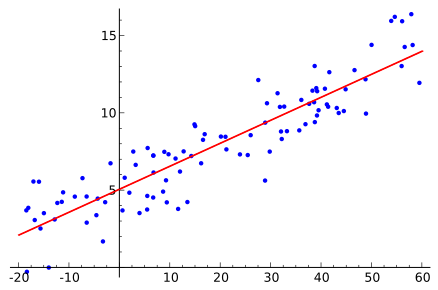
**1. Linear Regression**

Linear regression is one of the most basic types of regression in machine learning. The linear regression model consists of a predictor variable and a dependent variable related linearly to each other. In case the data involves more than one independent variable, then linear regression is called multiple linear regression models.

The below-given equation is used to denote the linear regression model:

y=mx+c+e

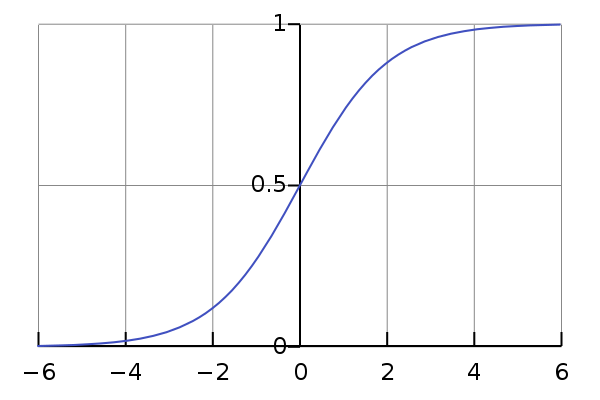
where m is the slope of the line, c is an intercept, and e represents the error in the model.



**2. Logistic Regression**

Logistic regression is one of the types of regression analysis technique, which gets used when the dependent variable is discrete. Example: 0 or 1, true or false, etc. This means the target variable can have only two values, and a sigmoid curve denotes the relation between the target variable and the independent variable.

Logit function is used in Logistic Regression to measure the relationship between the target variable and independent variables. Below is the equation that denotes the logistic regression.

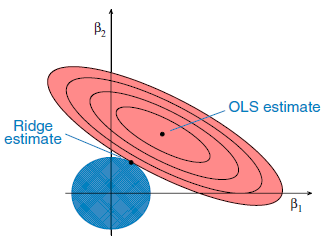


**3. Ridge Regression**

This is another one of the types of regression in machine learning which is usually used when there is a high correlation between the independent variables. This is because, in the case of multi collinear data, the least square estimates give unbiased values. But, in case the collinearity is very high, there can be some bias value. Therefore, a bias matrix is introduced in the equation of Ridge Regression. This is a powerful regression method where the model is less susceptible to overfitting.

Below is the equation used to denote the Ridge Regression, where the introduction of λ (lambda) solves the problem of multicollinearity:

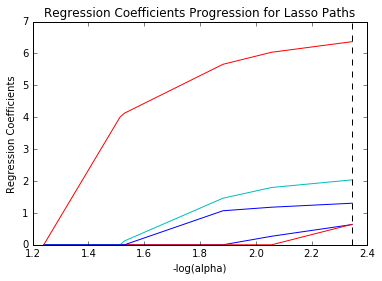
β = (X^{T}X + λ\*I)^{-1}X^{T}y



**4. Lasso Regression**

Lasso Regression is one of the types of regression in machine learning that performs regularization along with feature selection. It prohibits the absolute size of the regression coefficient. As a result, the coefficient value gets nearer to zero, which does not happen in the case of Ridge Regression.

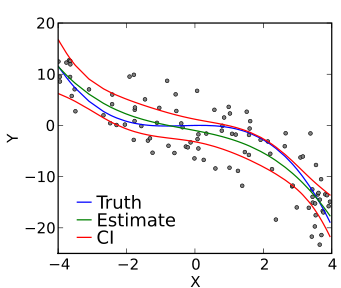
Due to this, feature selection gets used in Lasso Regression, which allows selecting a set of features from the dataset to build the model. In the case of Lasso Regression, only the required features are used, and the other ones are made zero. This helps in avoiding the overfitting in the model. In case the independent variables are highly collinear, then Lasso regression picks only one variable and makes other variables to shrink to zero.



**5. Polynomial Regression**

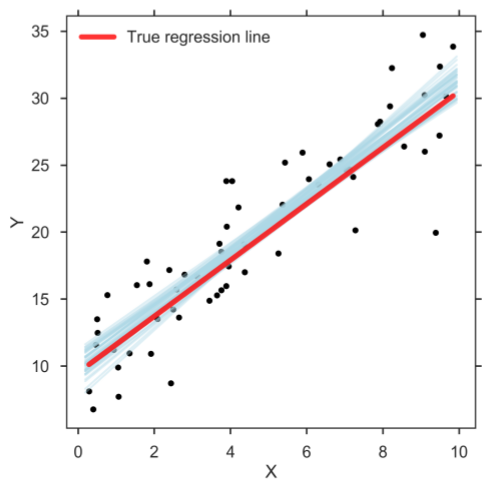
Polynomial Regression is another one of the types of regression analysis techniques in machine learning, which is the same as Multiple Linear Regression with a little modification. In Polynomial Regression, the relationship between independent and dependent variables, that is X and Y, is denoted by the n-th degree.

It is a linear model as an estimator. Least Mean Squared Method is used in Polynomial Regression also. The best fit line in Polynomial Regression that passes through all the data points is not a straight line, but a curved line, which depends upon the power of X or value of n.



**6. Bayesian Linear Regression**

Bayesian Regression is one of the types of regression in machine learning that uses the Bayes theorem to find out the value of regression coefficients. In this method of regression, the posterior distribution of the features is determined instead of finding the least-squares. Bayesian Linear Regression is like both Linear Regression and Ridge Regression but is more stable than the simple Linear Regression.



**PROBLEM STATEMENT**

In the given dataset, R&D Spend, Administration Cost and Marketing Spend of 50 Companies are given along with the profit earned. The target is to prepare an ML model which can predict the profit value of a company if the value of its R&D Spend, Administration Cost and Marketing Spend are given.

i) Construct Different Regression algorithms

ii) Divide the data into train set and test set

iii) Calculate different regression metrics

iv) Choose the best model

**CHAPTER 2**

**EXISTING METHODS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NAME OF PAPER NAME** | **SUBMITTED TO** | **YEAR OF PUBLICATION** | **NAME OF AUTHOR(S)** | **INFERENCE** |
| Startup Profit Predictor Using Machine Learning Techniques | LNNS | September 2022 | [Manasi Chhibber](https://link.springer.com/chapter/10.1007/978-981-19-4052-1_6#auth-Manasi-Chhibber) | The predictor makes use of four parameters, i.e., spend on R&D, administration, marketing and the location where the startup is based out of, and predicts an approximate value of the profit that it is most likely to make. On finding the perfect dataset, data preprocessing and data visualizations were done. The data was split for training and testing. The startup profit predictor was successfully built using random forest regressor model and it can be used for making profit predictions with 96.91% accuracy. |
| Predicting startup survival using first years financial statements | University of Zaragoza | August 2020 | Taylor and Francis | The models developed showed predictive capacity, but they did not reach that of the bankruptcy models made with mature companies. The analyzed period corresponded to a period of economic crisis. The study was repeated with data from another noncrisis period to enhance the validity of the results, and obtained similar results. |
| Predicting customer retention and profitability by using random forests and regression forests techniques | Empirical study | August 2005 | [Dirk Van den Poel](https://www.sciencedirect.com/science/article/abs/pii/S0957417405000965" \l "!) | demonstrate the benefits of analyzing different customer outcome variables simultaneously, since an extended investigation of the next buy–partial-defection–customer profitability triad indicates that one cannot fully understand a particular outcome without understanding the other related behavioral outcome variables. |
| Startups profit prediction using Multiple Linear Regression | IEEE | November 2021 | Aman Preet Gulati | Prediction of profit from the startup’s dataset with the features available to us. We’re using the 50-startups dataset for this problem statement and we will be using the concept of Multiple linear regression to predict the profit of startups companies. |

**CHAPTER 3**

**PROPOSED METHOD**

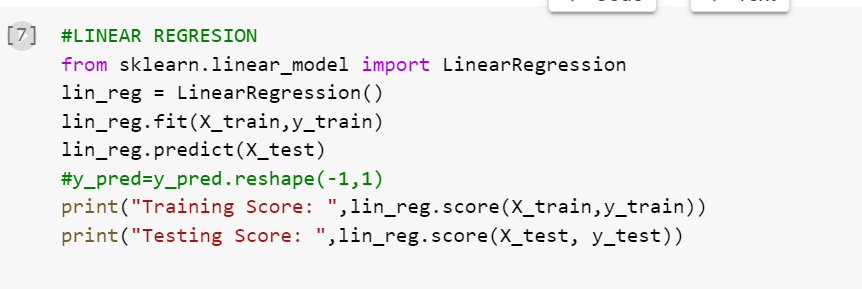
The steps involved in this project are given below:

* Firstly, the dataset is examined to get insights about what kind of data is present.
* Secondly, the dataset is split into train set and test set.
* Thirdly, the various regression models are fitted into the data.
* Fourthly, the scores for training and testing set is found out for each regression model
* Along with the scores , the RMSE and the cross val score is also found out for each regression model
* Finally, the best model is chosen based on the scores of the model

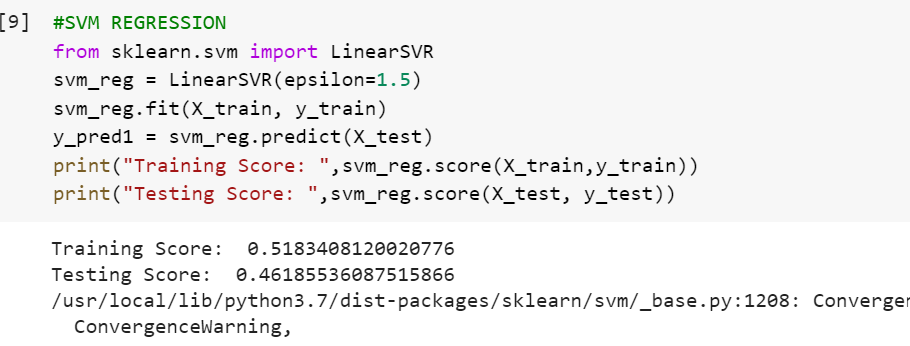
**CHAPTER 4**

**METHODOLOGY**

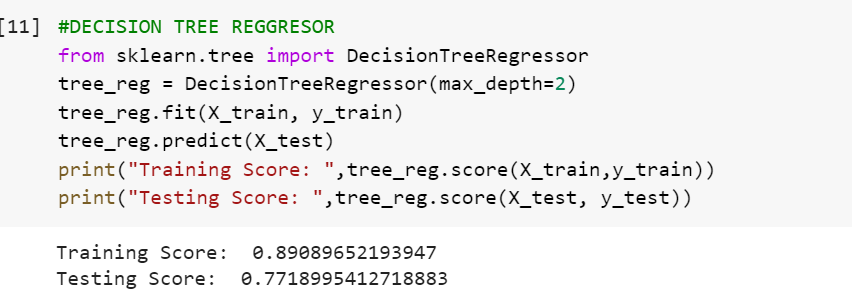
1st Model: The first model that has been implemented is a linear regression model. The code snippet for the Linear regressor is given below

****

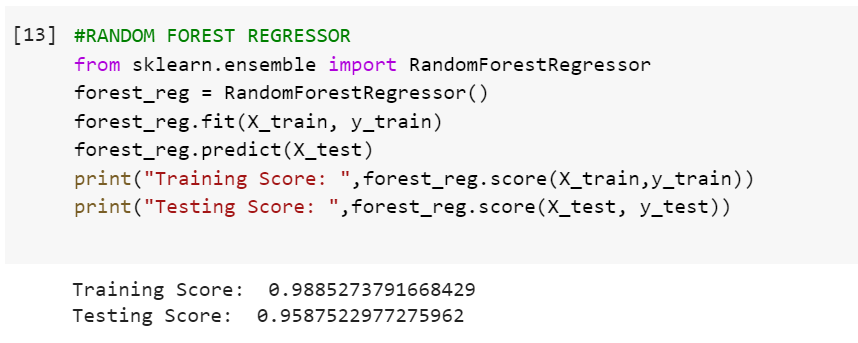
2nd Model: The second model that has been implemented is a SVM regression model. The code snippet for the SVM regressor is given below



3rd Model: The third model that has been implemented is a Decision Tree regression model. The code snippet for the Decision Tree regressor is given below



4th Model: The fourth model that has been implemented is a Random Forest regression model. The code snippet for the Random Forest regressor is given below

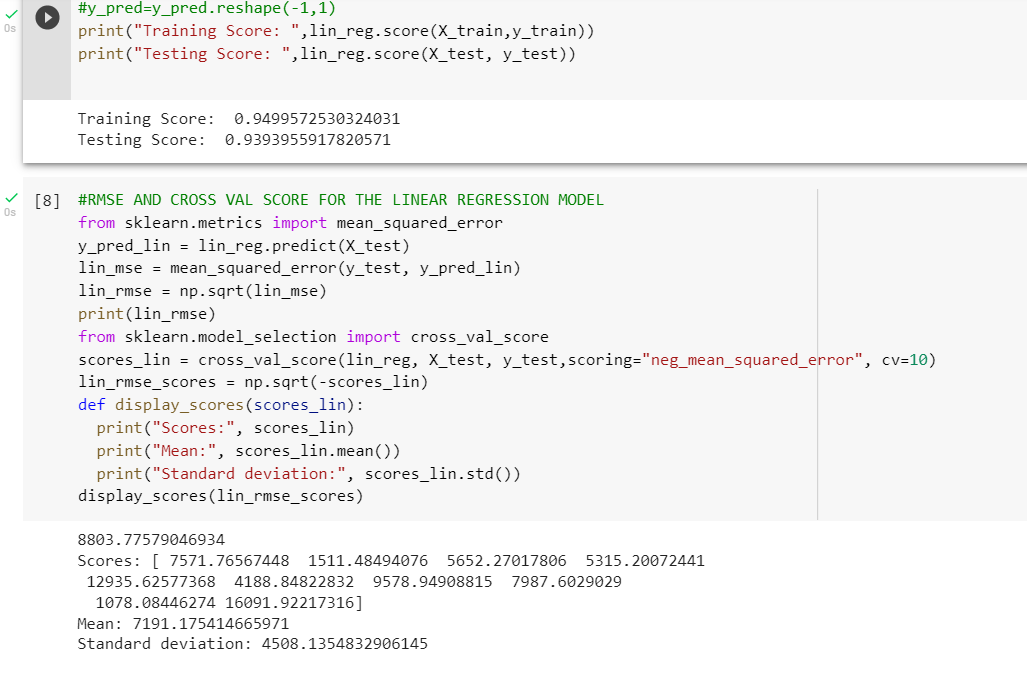


**CHAPTER 5**

**RESULTS**

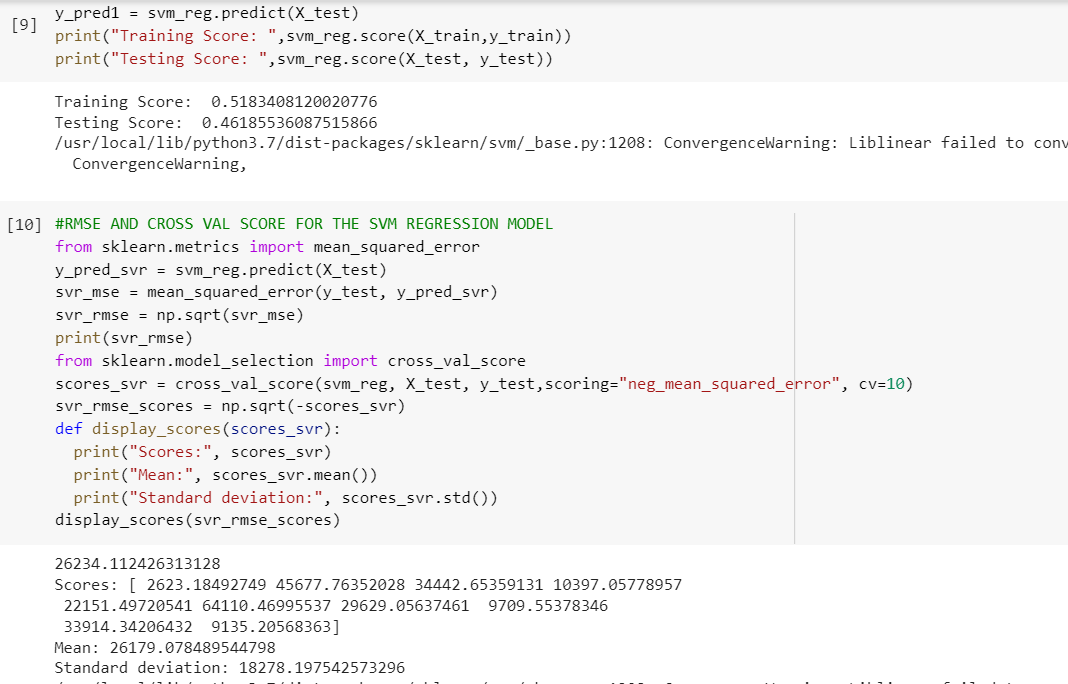
Linear Regression:

The training and the testing scores for a Linear Regression model are around 95% and 94% respectively. Along with the scores the Root mean square value and the cross val score for a linear regression model is found out and the code snippet for the same is given below.

****

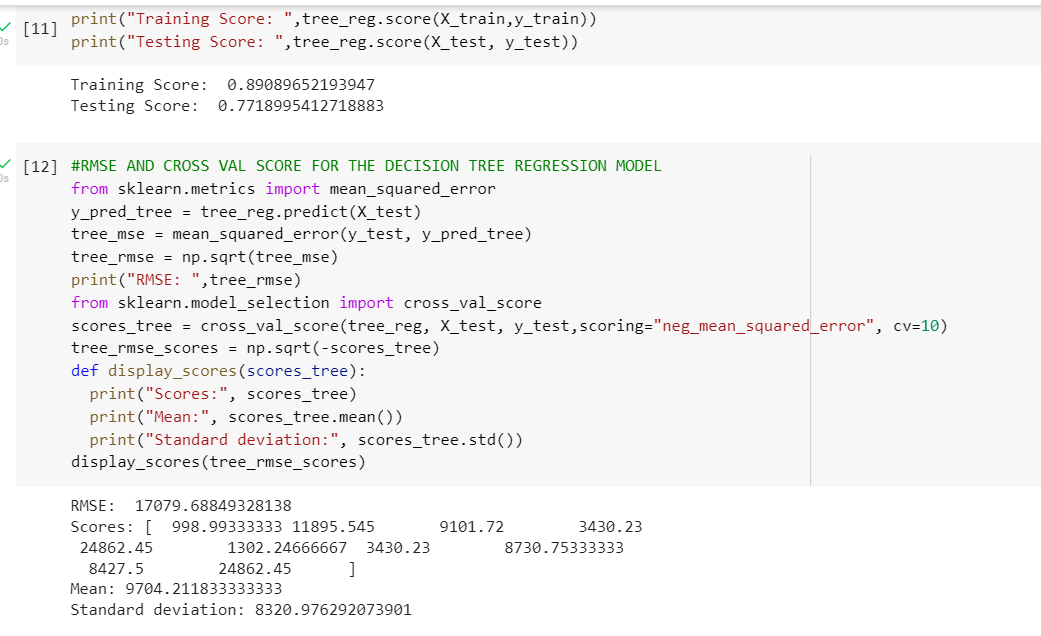
Support Vector Machine Regression:

The training and the testing scores for a SVM Regression model are around 52% and 46% respectively. Along with the scores the Root mean square value and the cross val score for a SVM regression model is found out and the code snippet for the same is given below.

****

Decision Tree Regression:

The training and the testing scores for a Decision Tree Regression model are around 89% and 77% respectively. Along with the scores the Root mean square value and the cross val score for a Decision Tree regression model is found out and the code snippet for the same is given below.

****

Random Forest Regression:

The training and the testing scores for a Random Forest Regression model are around 98% and 95% respectively. Along with the scores the Root mean square value and the cross val score for a Random Forest regression model is found out and the code snippet for the same is given below.

**CHAPTER 6**

**CONCLUSION**

The prediction of profit across 50 startups is found using four different regression model.

The four models are:

* Linear regression
* SVR
* Decision Tree regressor
* Random forest regressor

The best model is the Linear regressor as the score of the model is around 93% which is the best compared to the rest

**REFERENCES**

* <https://www.tandfonline.com/doi/abs/10.1080/00472778.2020.1750302>
* <https://www.analyticsvidhya.com/blog/2021/11/startups-profit-prediction-using-multiple-linear-regression/#:~:text=Multiple%20Linear%20Regression%20using%20Python&text=Table%20of%20Contents%20Introduction%20Working,Results%20Evaluate%20the%20model%20Plot%E2%80%A6>
* <https://www.sciencedirect.com/science/article/abs/pii/S0957417405000965>
* <https://en.wikipedia.org/wiki/Data_science>
* <https://medium.com/analytics-vidhya/linear-regression-in-python-predict-e-commerce-revenue-dd8e0c8bed2f>